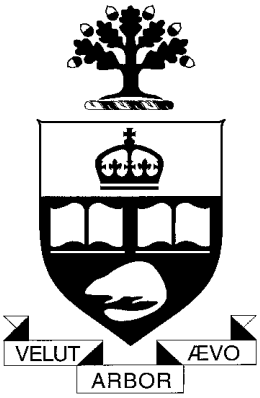

Weak Informativity and the Information in One Prior Relative to Another

by

Michael Evans
Department of Statistics
University of Toronto

and

Gun Ho Jang
Department of Statistics
University of Toronto

Technical Report No. 0907 December 3, 2009

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Weak Informativity and the Information in One Prior Relative to Another

Michael Evans and Gun Ho Jang
Department of Statistics
University of Toronto

Abstract

A question of some interest is how to characterize the amount of information that a prior puts into a statistical analysis. Rather than a general characterization of this information, we provide an approach to characterizing the amount of information a prior puts into an analysis, when compared to another base prior. The base prior is considered to be the prior that best reflects the current available information. Our purpose then, is to characterize priors that can be used as conservative inputs to an analysis, relative to the base prior, in the sense that they put less information into the analysis. The characterization that we provide is in terms of *a priori* measures of prior-data conflict.

1 Introduction

Suppose we have two proper priors Π_1 and Π_2 on a parameter space Θ for a statistical model $\{P_\theta : \theta \in \Theta\}$. A natural question to ask is: how do we compare the amount of information each of these priors puts into the problem? While there may seem to be natural intuitive ways to express this, such as prior variances, it seems difficult to characterize this precisely in general. For example, the consideration of several examples in Sections 3 and 4 makes it clear that using the variance of the prior is not appropriate for this task.

The motivation for this work comes from Gelman (2006) and Gelman *et al.* (2008), where the intuitively satisfying notion of weakly informative priors is introduced as a compromise between informative and noninformative priors. The basic idea is that we have a base prior Π_1, perhaps elicited, that we believe reflects our current information about θ, but we choose to be conservative in our inferences and select a prior Π_2 that puts less information into the analysis. While it is common to take Π_2 to be a noninformative prior, this can often produce difficulties when Π_2 is improper, and even when Π_2 is proper, it seems inappropriate as it completely discards the information we have about θ as expressed in Π_1.

To implement the idea of weak informativity we need a precise definition and so our purpose here is to give a definition of what it means for one prior to be weakly informative relative to another. We do this in Section 2 and it involves the notion of prior-data conflict. Intuitively, a prior-data conflict occurs when the prior places the bulk of its mass where the likelihood is relatively low. A completely noninformative prior should produce no prior-data conflicts *a priori*. Our definition of weak informativity can be expressed as saying that Π_2 is weakly informative relative to Π_1 whenever Π_2 produces fewer prior-data conflicts *a priori* than Π_1. This leads to a quantifiable expression of weak informativity that can be used to choose priors. In Section 3 we consider this definition in the context of several standard families of priors and it is seen to produce results that are intuitively reasonable. In Section 4 we consider applications of this concept in some data analysis problems. While our intuition about weak informativity is often borne out, we also find that in certain situations we have to be careful before calling a prior weakly informative.

First, however, we establish some notation and then review how we check for prior-data conflict. We suppose that $P_\theta(A) = \int_A f_\theta(x)\, \mu(dx)$, i.e., each P_θ is absolutely continuous with respect to a support measure μ on the sample space $\mathcal{X}$, with the density denoted by f_θ. With this formulation a prior Π leads to a prior predictive probability measure on $\mathcal{X}$ given by $M(A) = \int_\Theta P_\theta(A)\, \Pi(d\theta) = \int_A m(x)\, \mu(dx)$, where $m(x) = \int_\Theta f_\theta(x)\, \Pi(d\theta)$. If T is a minimal sufficient statistic for $\{P_\theta : \theta \in \Theta\}$, then it is well known that the posterior is the same whether we observe x or $T(x)$. So we will denote the posterior by $\Pi(\cdot \,|\, T)$ hereafter. Since T is minimal sufficient we know that the conditional distribution of x given T is independent of θ. We denote this conditional measure by $P(\cdot \,|\, T)$. The joint distribution $P_\theta \times \Pi$ can then be factored as

$$P_\theta \times \Pi = M \times \Pi(\cdot \,|\, x) = P(\cdot \,|\, T) \times M_T \times \Pi(\cdot \,|\, T) \tag{1}$$

where M_T is the marginal prior predictive distribution of T.

While much of Bayesian analysis focuses on the third factor in (1), there are also roles in a statistical analysis for $P(\cdot \,|\, T)$ and M_T. As discussed in Evans and Moshonov (2006, 2007), $P(\cdot \,|\, T)$ is available for checking the sampling model, e.g., if x is a surprising value from this distribution, then we have evidence that the model $\{P_\theta : \theta \in \Theta\}$ is incorrect. Further it is argued that, if we conclude that we have no evidence against the model, then the factor M_T is available for checking whether or not there is any prior-data conflict. So if $T(x)$ is a surprising value from M_T, then we have evidence that the prior Π is placing most of its mass on θ values where the likelihood is relatively low. This is supported by the fact that T is equivalent to the likelihood map. Finally, if we have no evidence against the model, and no evidence of prior-data conflict, then $\Pi(\cdot \,|\, T)$ is available for probability statements about θ. Actually the issues involved in model checking and checking for prior-data conflict are more involved than this (see, for example, the cited references and Section 5), but (1) gives the basic idea that the full information, as expressed by the joint distribution of (θ, x), splits into components, each of which is available for a specific purpose in a statistical analysis.

Accordingly we restrict ourselves here, for any discussions about the respective merits of priors, to working with M_T. One issue that needs to be addressed is how one is to compare the observed data x_0 to $P(\cdot \mid T)$ or compare $t_0 = T(x_0)$ to M_T. In essence we need a measure of surprise. Perhaps the best measure of surprise is the P-value. Effectively, we are in the situation where we have a value from a single fixed distribution and we need to specify the appropriate P-value to use. In Evans and Moshonov (2006, 2007) the P-value for checking for prior-data conflict was based on the prior predictive density m_T, namely,

$$M_T(m_T(t) \leq m_T(t_0)). \tag{2}$$

A difficulty with (2) is that it is not generally invariant to the choice of the minimal sufficient statistic T. A general invariant P-value is developed in Evans and Jang (2008) for such situations. When applied to (2), this leads to using the invariant P-value

$$M_T(m_T^*(t) \leq m_T^*(t_0)) \tag{3}$$

instead, where $m_T^*(t) = \int_\Theta f_{\theta T}^*(t)\,\Pi(d\theta)$, $f_{\theta T}^*(t) = \int_{T^{-1}\{t\}} f_\theta(x)\,\nu_{T^{-1}\{t\}}(dx)$ and $\nu_{T^{-1}\{t\}}$ is volume measure on $T^{-1}\{t\}$. Note that it can be shown that the marginal density of T, with respect to volume measure on the range space for T, is given by $f_{\theta T}(t) = \int_{T^{-1}\{t\}} f_\theta(x) J_T(x)\,\nu_{T^{-1}\{t\}}(dx)$ where $J_T(x) = (\det(dT(x) \circ dT'(x)))^{-1/2}$ and dT is the differential of T. So $f_{\theta T}^*(t)$ is an adjustment of $f_{\theta T}(t)$ where we do not allow the volume distortions induced by T to affect the density. It is also shown in Evans and Jang (2008) that $f_{\theta T}^*(t) = f_{\theta,T}(t) E_\theta(J_T(X)^{-1} \mid T = t)$ so $f_{\theta T}^*(t)$ can be thought of as a multiplication of the usual marginal density by the expected volume correction given $T = t$. We will refer to the invariant P-value (3) throughout the remainder of our discussion but note that, for many of the examples in this paper, $J_T(x)$ is constant, e.g., whenever T is linear in x, and then (2) and (3) are equal.

2 Comparing Priors

There are a variety of measures of information used in statistics. Several measures have been based on the concept of entropy, e.g., see Lindley (1956) and Bernardo (1979). While these measures have their virtues, we note that their coding theory interpretations can seem somewhat abstract in statistical contexts and they can suffer from nonexistence in certain problems. Also, Kass and Wasserman (1995) contains some discussion concerned with expressing the absolute information content of a prior in terms of additional sample values. Rather than adopting these approaches, we consider here comparing priors based on their tendencies to produce prior-data conflicts. The basic intuitive idea is that a prior which produces fewer prior-data conflicts than another, is putting less information into an analysis. This formulation of the relative amount of information put into an analysis, has a direct interpretation in terms of statistical consequences.

Suppose that an analyst has in mind a prior Π_1 that they believe represents the information at hand concerning θ. The analyst, however, prefers to use a prior Π_2 that is somewhat conservative, with respect to the amount of information put into the analysis, when compared to Π_1. The motivation for this lies with a desire not to put too much information into the analysis via the prior, and avoid the use of completely noninformative priors. Often priors that are characterized as being noninformative are improper and their use can be challenged for a variety of reasons. In such a situation it seems reasonable to consider Π_1 as a base prior and then compare all other priors to it. This idea comes from Gelman (2006) and leads to the notion of weakly informative priors, but without a precise characterization of this concept.

For a given prior Π_1 and observed value $t_0 = T(x)$ then, from (3), we have that $M_{1T}(m^*_{1T}(t) \leq m^*_{1T}(t_0))$ is the relevant quantity for assessing whether or not there is prior-data conflict with Π_1. Before we observe data, however, we have no way of knowing if we will have a prior-data conflict. Accordingly, since the analyst has determined that Π_1 best reflects the available information, it is reasonable to consider the prior distribution of $P_1(t_0) = M_{1T}(m^*_{1T}(t) \leq m^*_{1T}(t_0))$ when $t_0 \sim M_{1T}$. Of course, this is effectively uniformly distributed (exactly so when $m^*_{1T}(t)$ has a continuous distribution when $t \sim M_{1T}$) and this expresses the fact that all the information about assessing whether or not a prior-data conflict exists, is contained in the P-value, with no need to compare the P-value to its distribution.

Consider now, however, the distribution of $P_2(t_0) = M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$ which is used to check whether or not there is prior-data conflict with respect to Π_2. Given we have identified that *a priori* the appropriate distribution of t_0 is M_{1T}, at least for inferences about an unobserved value, then $P_2(t_0)$ is not uniformly distributed. In fact, from the distribution of $P_2(t_0)$ we can obtain an intuitively reasonable idea of what it means for a prior Π_2 to be weakly informative relative to Π_1. For suppose that the prior distribution of $P_2(t_0)$ clusters around 1. This implies that, if we were to use Π_2 as the prior when Π_1 is appropriate, then there is a small prior probability that a prior-data conflict would arise. Similarly, if the prior distribution of $P_2(t_0)$ clusters around 0, then there is a large prior probability that a prior-data conflict would arise. If one prior distribution results in a larger prior probability of there being a prior-data conflict than another, then it seems reasonable to say that the first prior is more informative than the second. In fact, a completely noninformative prior should never produce prior-data conflicts.

So we compare the distribution of $P_2(t_0)$ when $t_0 \sim M_{1T}$, to the distribution of $P_1(t_0)$ when $t_0 \sim M_{1T}$, and do this in a way that is relevant to the prior probability of obtaining a prior-data conflict. One approach to this comparison is to select a γ-quantile $x_\gamma \in [0, 1]$ of the distribution of $P_1(t_0)$, and then compute the probability

$$M_{1T}(P_2(t_0) \leq x_\gamma). \tag{4}$$

The value γ is presumably some cut-off, dependent on the application, where we will consider that evidence of a prior-data conflict exists whenever $P_1(t_0) \leq \gamma$.

Of course, if $m^*_{1T}(t_0)$ has a continuous distribution when $t_0 \sim M_{1T}$, then $x_\gamma = \gamma$. Our criterion for the weak informativity of Π_2 relative to Π_1 will then be that (4) is less than or equal to x_γ. This indicates that the prior probability of obtaining a prior-data conflict under Π_2 is no greater than when Π_1 is used, at least when we have identified Π_1 as our correct prior.

Definition. If (4) is less than or equal to x_γ, then Π_2 is *weakly informative relative to* Π_1 *at level* γ. If Π_2 is weakly informative relative to Π_1 at level γ for every $\gamma \leq \gamma_0$, then Π_2 is *uniformly weakly informative relative to* Π_1 *at level* γ_0. If Π_2 is weakly informative relative to Π_1 at level γ for every γ, then Π_2 is *uniformly weakly informative relative to* Π_1.

While it would be appealing to be able to choose a prior Π_2 that is uniformly weakly informative relative to Π_1, this may not always be preferable. The criterion of uniformly weakly informative at level γ_0 seems much more generally applicable.

While (4) seems difficult to work with, the following result is proved in the Appendix and gives a simpler expression.

Lemma 1. Suppose $P_i(t)$ has a continuous distribution under M_{iT} for $i = 1, 2$. Then there exists r_γ such that $M_{1T}(P_2(t) \leq \gamma) = M_{1T}(m^*_{2T}(t) \leq r_\gamma)$, and Π_2 is weakly informative at level γ relative to Π_1 whenever $M_{1T}(m^*_{2T}(t) \leq r_\gamma) \leq \gamma$. Furthermore, Π_2 is uniformly weakly informative relative to Π_1 if and only if $M_{1T}(m^*_{2T}(t) \leq m^*_{2T}(t_0)) \leq M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$ for every t_0.

Lemma 1 typically applies when we are dealing with continuous distributions on $\mathcal{X}$. It can also be shown that $P_i(t)$ has a continuous distribution under M_{iT} if and only if $m^*_{iT}(t)$ has a continuous distribution under M_{iT}.

Once we have selected γ, the degree of weak informativity of a prior Π_2 relative to Π_1 can be assessed by comparing $M_{1T}(P_2(t_0) \leq x_\gamma)$ to x_γ via the ratio

$$1 - M_{1T}(P_2(t_0) \leq x_\gamma)/x_\gamma. \tag{5}$$

If Π_2 is weakly informative relative to Π_1 at level γ, then (5) tells us the proportion of fewer prior-data conflicts we can expect *a priori* when using Π_2 rather than Π_1. Thus (5) provides a measure of how much less informative Π_2 is than Π_1, e.g., we might ask for a prior Π_2 such that (5) equals 50%.

As we will see in the examples, it makes sense to talk of one prior being *asymptotically weakly informative at level* γ with respect to another prior, whenever (4) is bounded above by γ in the limit as the amount of data increases. In several cases this simplifies matters considerably as an asymptotically weakly informative prior is easy to find and may still be weakly informative for finite amounts of data.

3 Deriving Weakly Informative Priors

We consider several examples of priors that arise in applications. These examples support our definition of weak informativity and also lead to some insights

into choosing priors.

3.1 Comparing Normal Priors

Suppose we have a sample $x = (x_1, \ldots, x_n)$ from a $N(\mu, 1)$ distribution where μ is unknown. Then $t = T(x) = \bar{x} \sim N(\mu, 1/n)$ is minimal sufficient and since T is linear there is constant volume distortion and so this can be ignored. Suppose that the prior Π_1 on μ is a $N(\mu_0, \sigma_1^2)$ distribution with μ_0 and σ_1^2 known. We then have that M_{1T} is the $N(\mu_0, 1/n + \sigma_1^2)$ distribution. Now suppose that Π_2 is a $N(\mu_0, \sigma_2^2)$ distribution with σ_2^2 known. Then M_{2T} is the $N(\mu_0, 1/n + \sigma_2^2)$ distribution and

$$\begin{aligned} P_2(t_0) &= M_{2T}(m_{2T}^*(t) \leq m_{2T}^*(t_0)) = M_{2T}(m_{2T}(t) \leq m_{2T}(t_0)) \\ &= M_{2T}((t-\mu_0)^2 \geq (t_0-\mu_0)^2) = 1 - G_1((t_0-\mu_0)^2/(1/n+\sigma_2^2)), \end{aligned}$$

where G_k denotes the Chi-squared(k) distribution function. Now under M_{1T} we have that $(t_0-\mu_0)^2/(1/n+\sigma_1^2) \sim$ Chi-squared(1). Therefore,

$$\begin{aligned} M_{1T}(P_2(t_0) \leq \gamma) &= M_{1T}(1 - G_1((t_0-\mu_0)^2/(1/n+\sigma_2^2)) \leq \gamma) \\ &= M_{1T}\left(\frac{(t_0-\mu_0)^2}{1/n+\sigma_1^2} \geq \frac{1/n+\sigma_2^2}{1/n+\sigma_1^2} G_1^{-1}(1-\gamma)\right) \\ &= 1 - G_1\left(\frac{1/n+\sigma_2^2}{1/n+\sigma_1^2} G_1^{-1}(1-\gamma)\right). \end{aligned} \tag{6}$$

We see immediately that (6) will be less than γ if and only if $\sigma_2 > \sigma_1$. In other words Π_2 will be uniformly weakly informative relative to Π_1 if and only if Π_2 is more diffuse than Π_1. Note that $M_{1T}(P_2(t_0) \leq \gamma)$ converges to 0 as $\sigma_2^2 \to \infty$ to reflect noninformativity. Also, as $n \to \infty$, then (6) increases to $1 - G_1((\sigma_2^2/\sigma_1^2) G_1^{-1}(1-\gamma))$ and so we could ignore n and choose σ_2^2 conservatively based on this limit, to obtain an asymptotically uniformly weakly informative prior, as we know this value of σ_2^2 will also be weakly informative for finite n.

If we specify that we want (5) to equal $p \in [0,1]$, then (6) implies that $\sigma_2^2 = (1/n+\sigma_1^2)(G_1^{-1}(1-\gamma+p\gamma)/G_1^{-1}(1-\gamma)) - 1/n$. Such a choice will give a proportion p fewer prior-data conflicts at level γ than the base prior. This decreases to $\sigma_1^2 G_1^{-1}(1-\gamma+p\gamma)/G_1^{-1}(1-\gamma)$ as $n \to \infty$ and so the more data we have the less extra variance we need for Π_2 for weak informativity.

We can generalize this to $t \sim N_k(\mu, n^{-1}I)$ with Π_1 given by $\mu \sim N_k(\mu_0, \Sigma_1)$. Note we have that M_{iT} is the $N_k(\mu_0, n^{-1}I + \Sigma_i)$ distribution. It is then easy to see that $P_2(t_0) = 1 - G_k((t_0-\mu_0)'(n^{-1}I+\Sigma_2)^{-1}(t_0-\mu_0))$ and

$$M_{1T}(P_2(t_0) \leq \gamma) = M_{1T}((t_0-\mu_0)'(n^{-1}I+\Sigma_2)^{-1}(t_0-\mu_0) \geq G_k^{-1}(1-\gamma)). \tag{7}$$

Note that (7) increases to the probability that $(t_0-\mu_0)'{\Sigma_2}^{-1}(t_0-\mu_0) \geq G_k^{-1}(1-\gamma)$, when $t_0 \sim N_k(\mu_0, \Sigma_1)$, as $n \to \infty$. This probability can be easily computed via simulation.

The following result is proved in the Appendix.

Theorem 1. For a sample of n from the statistical model $\{N_k(\mu, I) : \mu \in R^k\}$, a $N_k(\mu_0, \Sigma_2)$ prior is uniformly weakly informative relative to a $N_k(\mu_0, \Sigma_1)$ prior if and only if $\Sigma_2 - \Sigma_1$ is positive semidefinite.

The necessary part of Theorem 1 is much more difficult than the $k = 1$ case and shows that we cannot have a $N_k(\mu_0, \Sigma_2)$ prior uniformly weakly informative relative to a $N_k(\mu_0, \Sigma_1)$ prior unless $\Sigma_2 \geq \Sigma_1$. For the choice of Σ_2 we have that, if Σ_1 and Σ_2 are arbitrary $k \times k$ positive definite matrices, then $r\Sigma_2 \geq \Sigma_1$ whenever $r \geq \lambda_k(\Sigma_2)/\lambda_1(\Sigma_2)$ where $\lambda_i(\Sigma)$ denotes the i-th ordered eigenvalue of Σ. Also, if $\Sigma_i = QD_iQ'$ is the spectral decomposition of Σ_i, then $\Sigma_2 \geq \Sigma_1$ whenever $\lambda_i(\Sigma_2) \geq \lambda_i(\Sigma_1)$ for $i = 1, \ldots, k$.

3.2 Comparing a t Prior with a Normal Prior

It is not uncommon to find t priors being substituted for normal priors on location parameters. Suppose $x = (x_1, \ldots, x_n)$ is a sample from a $N(\mu, 1)$ distribution where μ is unknown. We take Π_1 to be a $N(\mu_0, \sigma_1^2)$ distribution and Π_2 to be a $t_1(\mu_0, \sigma_2^2, \lambda)$ distribution, i.e., $t_1(\mu_0, \sigma_2^2, \lambda)$ denotes the distribution of $\mu_0 + \sigma_2 z$ with z distributed as a 1-dimensional t distribution with λ degrees of freedom. We then want to determine σ_2^2 and λ so that the $t_1(\mu_0, \sigma_2^2, \lambda)$ prior is weakly informative relative to the normal prior.

We consider first the limiting case as $n \to \infty$. The limiting prior predictive distribution of the minimal sufficient statistic $T(x) = \bar{x}$ is $N(\mu_0, \sigma_1^2)$ while $P_2(t_0)$ converges in distribution to $1 - H_{1,\lambda}((t_0 - \mu_0)^2/\sigma_2^2)$ where $H_{1,\lambda}$ is the distribution function of an $F_{1,\lambda}$ distribution. This implies that (4) converges to $1 - G_1((\sigma_2^2/\sigma_1^2)H_{1,\lambda}^{-1}(1-\gamma))$ and this is less than or equal to γ if and only if $\sigma_2^2/\sigma_1^2 \geq G_1^{-1}(1-\gamma)/H_{1,\lambda}^{-1}(1-\gamma))$. So to have that Π_2 is asymptotically weakly informative relative to Π_1 at level γ we must choose σ_2^2 large enough. Clearly we have that Π_2 is asymptotically uniformly weakly informative relative to Π_1 if and only if

$$\sigma_2^2/\sigma_1^2 \geq \sup_{\gamma \in [0,1]} G_1^{-1}(1-\gamma)/H_{1,\lambda}^{-1}(1-\gamma).$$

For example, consider the following table.

λ	0.5	1	3	100
$\sup_{\gamma\in[0,1]} G_1^{-1}(1-\gamma)/H_{1,\lambda}^{-1}(1-\gamma))$	0.4569	0.6366	0.8488	0.9950

We see that for a Cauchy prior we need to have $\sigma_2^2 \geq \sigma_1^2(0.6366)$ for this prior to be uniformly weakly informative with respect to a $N(\mu_0, \sigma_1^2)$ prior. When we have a $t_1(\mu_0, \sigma_2^2, 3)$ prior, then this has variance $3\sigma_2^2$ and, if we choose σ_2^2 so that this prior also has variance σ_1^2, then $\sigma_2^2/\sigma_1^2 = 1/3$ and this is less than 0.8488 and so is not uniformly weakly informative. So a $t_1(\mu_0, \sigma_2^2, 3)$ prior has to have variance at least equal to $(2.5464)\sigma_1^2$ if we want it to be uniformly weakly informative relative to a $N(\mu_0, \sigma_1^2)$ prior. This is somewhat surprising and undoubtedly is caused by the peakedness of the t distribution. Note that

$\sup_{\gamma\in[0,1]} G_1^{-1}(1-\gamma)/H_{1,\lambda}^{-1}(1-\gamma)) \to 1$ as $\lambda \to \infty$ so this increase in variance, for the t prior over the normal prior, decreases as we increase the degrees of freedom.

The situation for finite n is covered by the following result proved in the Appendix.

Theorem 2. For a sample of n from the statistical model $\{N(\mu,1) : \mu \in R^1\}$, a $t_1(\mu_0, \sigma_2^2, \lambda)$ prior is uniformly weakly informative relative to a $N_1(\mu_0, \sigma_1^2)$ prior whenever $\sigma_2^2 \geq \sigma_{0n}^2$, where σ_{0n}^2 is the unique solution of $(1/n + \sigma_1^2)^{-1/2} = \int_0^\infty (1/n + \sigma_{0n}^2/u)^{-1/2} k_\lambda(u)\, du$ with k_λ the $\text{Gamma}_{\text{rate}}(\lambda/2, \lambda/2)$ density. Further, σ_{0n}^2/σ_1^2 increases to

$$\sup_{\gamma\in[0,1]} \frac{G_1^{-1}(1-\gamma)}{H_{1,\lambda}^{-1}(1-\gamma)} = \frac{2}{\lambda}\frac{\Gamma^2((\lambda+1)/2)}{\Gamma^2(\lambda/2)} \tag{8}$$

as $n \to \infty$ and so a $t_1(\mu_0, \sigma_2^2, \lambda)$ prior is asymptotically uniformly weakly informative if and only if σ_2^2/σ_1^2 is greater than or equal to (8).

Theorem 2 establishes that we can conservatively use (8) to select a uniformly weakly informative t prior.

In Figure 1 we have plotted the value of (4), that arises with various $t_1(0, \sigma_2^2, 3)$ priors where σ_2^2 is chosen in a variety of ways together with the 45-degree line. A uniformly weakly informative prior will have (4) always below the 45-degree line, while uniformly weakly informative prior at level γ_0 will have (4) below the 45-degree line to the left of γ_0 and possibly above to the right of γ_0. For example, when $\sigma_2^2 = 1/3$ then the $t_1(0, \sigma_2^2, 3)$ prior and the $N(0,1)$ prior have the same variance. We see that this prior is only uniformly weakly informative at level $\gamma_0 = 0.0357$ and is not uniformly weakly informative.

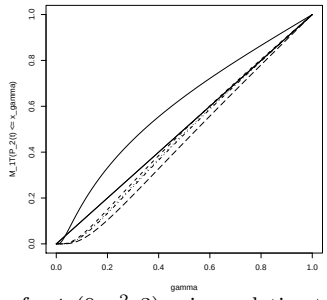

Figure 1: Plot of (4) versus γ for $t_1(0, \sigma_2^2, 3)$ priors relative to a $N(0,1)$ prior when $n = 20$ where σ_2^2 is chosen to match variances (thick solid line), match the MAD (dashed line), just achieve uniform weak informativity (dotted line), just achieve asymptotic uniform weak informativity (dash-dot line), and equal to 1 (long-dashed line).

Note that (5) converges to $1 - G_1((\sigma_2^2/\sigma_1^2)H_{1,\lambda}^{-1}(1-\gamma))/\gamma$ as $n \to \infty$, and setting this equal to p, implies that $\sigma_2^2 = \sigma_1^2 G_1^{-1}(1-\gamma+\gamma p)/H_{1,\lambda}^{-1}(1-\gamma)$ which converges, as $\lambda \to \infty$, to the result we obtained in Section 3.1. So when $\lambda = 3, \gamma = 0.05$ and $p = 0.5$ we must have $\sigma_2^2/\sigma_1^2 = 5.0239/10.1280 = 0.49604$.

Our analysis indicates that one has to be careful about the scaling of the t prior if we want to say that the t prior is less informative than a normal prior,

at least when we want uniform weak informativity. This is undoubtedly due to the peakedness of the t prior.

Consider now comparing a multivariate t prior to a multivariate normal prior. Let $t_k(\mu_0, \Sigma_2, \lambda)$ denote the k-dimensional t distribution given by $\mu_0 + \Sigma_2^{1/2} z$, where $\Sigma_2^{1/2}$ is a square root of the positive definite matrix Σ_2 and z has a k-dimensional t distribution with λ degrees of freedom. This is somewhat more complicated than the normal case but we have proved the following result in the Appendix which provides sufficient conditions for the asymptotic uniform weak informativity.

Theorem 3. When sampling from the statistical model $\{N_k(\mu, I) : \mu \in R^k\}$, a $t_k(\mu_0, \Sigma_2, \lambda)$ prior is asymptotically uniformly weakly informative relative to a $N_k(\mu_0, \Sigma_1)$ prior whenever $\Sigma_2 - \tau_\lambda^2 \Sigma_1$ is positive semidefinite, where $\tau_\lambda^2 = (2/\lambda)\Gamma^{2/k}((k+\lambda)/2)/\Gamma^{2/k}(\lambda/2)$.

For the choice of Σ_2 we have that, if Σ_1 and Σ_2 are arbitrary $k \times k$ positive definite matrices, then $r\Sigma_2 \geq \tau_\lambda^2 \Sigma_1$ whenever $r \geq \tau_\lambda^2 \lambda_k(\Sigma_2)/\lambda_1(\Sigma_2)$ where $\lambda_i(\Sigma)$ denotes the i-th ordered eigenvalue of Σ. Also, if $\Sigma_i = QD_iQ'$ is the spectral decomposition of Σ_i, then $\Sigma_2 \geq \Sigma_1$ whenever $\lambda_i(\Sigma_2) \geq \tau_\lambda^2 \lambda_i(\Sigma_1)$ for $i = 1, \ldots, k$.

3.3 Comparing Inverse Gamma Priors

Suppose now that we have a sample $x = (x_1, \ldots, x_n)$ from a $N(0, \sigma^2)$ distribution where σ^2 is unknown. Then $t = T(x) = (x_1^2 + \cdots + x_n^2)/n$ is minimal sufficient and $T \sim \text{Gamma}_{\text{rate}}(n/2, n/2\sigma^2)$. Now suppose that we take Π_i to be an inverse gamma prior on σ^2, namely, $\sigma^{-2} \sim \text{Gamma}_{\text{rate}}(\alpha_i, \beta_i)$. From this we get that $\alpha_i T/\beta_i \sim F(n, 2\alpha_i)$ and, since $J_T(x) = (4x'x/n)^{-1/2} = (4t/n)^{-1/2}, m_{iT,n}^*(t) = m_{iT,n}(t)(4t/n)^{1/2} \propto t^{(n-1)/2}(1 + nt/2\beta_i)^{-n/2-\alpha_i}$, which implies

$$P_{i,n}(t_0) = M_{iT,n}(t^{(n-1)/2}(1 + nt/2\beta_i)^{-n/2-\alpha_i} \leq t_0^{(n-1)/2}(1 + nt_0/2\beta_i)^{-n/2-\alpha_i}).$$

We want to investigate the weak informativity of a $\text{Gamma}_{\text{rate}}(\alpha_2, \beta_2)$ prior relative to a $\text{Gamma}_{\text{rate}}(\alpha_1, \beta_1)$ prior. For finite n this is a very difficult problem so we simplify this by considering only the asymptotic case. When the prior is Π_i, then, as $n \to \infty$, we have that $m_{iT,n}(t) \to m_{iT}(t) = (\beta_i^{\alpha_i}/\Gamma(\alpha_i))t^{-\alpha_i - 1}e^{-\beta_i/t}$, i.e., $1/t \sim \text{Gamma}_{\text{rate}}(\alpha_i, \beta_i)$ in the limit. Therefore, $P_{2,n}(t_0) \to P_2(t_0) = \Pi_2(t^{-\alpha_2 - 1/2}e^{-\beta_2/t} \leq t_0^{-\alpha_2 - 1/2}e^{-\beta_2/t_0})$ and we want to determine conditions on (α_2, β_2) so that $\Pi_1(P_2(t) \leq \gamma) \leq \gamma$.

While results can be obtained for this problem it is still rather difficult. It is greatly simplified, however, if we impose a natural restriction on (α_2, β_2). In particular, we want the location of the bulk of the mass for Π_2 to be located roughly in the same place as the bulk of the mass for Π_1. Accordingly, we could require the priors to have the same means or modes but, as it turns out, the constraint that requires the modes of the m_{iT}^* functions to be the same greatly simplifies the analysis. Actually $m_{iT,n}^*(t)$ converges to 0 but the n's cancel in the inequalities defining $P_{i,n}(t_0)$ and so we can define $m_{iT,n}^*(t) =$

$t^{-\alpha_i-1/2}e^{-\beta_i/t}$ which has its mode at $t = \beta_i/(\alpha_i + 1/2)$. Therefore, we must have $\beta_2/(\alpha_2 + 1/2) = \beta_1/(\alpha_1 + 1/2)$ so that (α_2, β_2) lies on the line through the points $(0, \beta_1/2(\alpha_1 + 1/2))$ and (α_1, β_1). We prove the following result in the Appendix.

Theorem 4. Suppose we use a $\text{Gamma}_{\text{rate}}(\alpha_1, \beta_1)$ prior on $1/\sigma^2$ when sampling from the statistical model $\{N(0, \sigma^2) : \sigma^2 > 0\}$. Then a $\text{Gamma}_{\text{rate}}(\alpha_2, \beta_2)$ prior on $1/\sigma^2$, with $\beta_2/(\alpha_2 + 1/2) = \beta_1/(\alpha_1 + 1/2)$, is asymptotically weakly informative relative to the $\text{Gamma}_{\text{rate}}(\alpha_1, \beta_1)$ prior whenever $\alpha_2 \leq \alpha_1$ and $\beta_2 = \beta_1(\alpha_2 + 1/2)/(\alpha_1 + 1/2)$ or, equivalently, whenever $\beta_1/2(\alpha_1 + 1/2) \leq \beta_2 \leq \beta_1$ and $\alpha_2 = (\alpha_1 + 1/2)\beta_2/\beta_1 - 1/2$.

Of particular interest here is that we cannot reduce the rate parameter β_2 arbitrarily close to 0 and be guaranteed asymptotic weak informativity.

4 Applications

We consider now some applications of determining weakly informative priors.

4.1 Weakly Informative Beta Priors for the Binomial

Suppose that $T \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$. This implies that $m_T(t) = \binom{n}{t}\Gamma(\alpha + \beta)\Gamma(t + \alpha)\Gamma(n - t + \beta)/\Gamma(\alpha)\Gamma(\beta)\Gamma(n + \alpha + \beta)$ and from this we can compute (4) for various choices of (α, β).

As a specific example, suppose that $n = 20$, the base prior is given by $(\alpha, \beta) = (6, 6)$, and we take $\gamma = 0.05$ so that $x_{0.05} = 0.0588$. As alternatives to this base prior, we consider $\text{Beta}(\alpha, \beta)$ priors. In Figure 2 we have plotted all the (α, β) corresponding to $\text{Beta}(\alpha, \beta)$ distributions that are weakly informative with respect to the $\text{Beta}(6, 6)$ distribution at level 0.05, together with the subset of all (α, β) corresponding to $\text{Beta}(\alpha, \beta)$ distributions that are uniformly weakly informative relative to the $\text{Beta}(6, 6)$ distribution. The graph on the left corresponds to $n = 20$, the middle graph corresponds to $n = 100$, and the graph on the right corresponds to $n = \infty$. The plot for $n = 20$ shows some anomalous effects due to the discreteness of the prior predictive distributions and these effects disappear as n increases. In such an application we may choose to restrict to symmetric priors as this fixes the primary location of the prior mass. For example, when $n = 20$, a $\text{Beta}(\alpha, \alpha)$ prior for α satisfying $1 \leq \alpha \leq 12.3639$ is uniformly weakly informative with respect to the $\text{Beta}(6, 6)$ prior and we see that values of $\alpha > 6$ are eliminated as n increases.

4.2 Weakly Informative Priors for a Location-Scale Model

Suppose that $x = (x_1, \ldots, x_n)$ is a sample from a $N(\mu, \sigma^2)$ distribution where $\mu \in R^1$ and $\sigma^2 > 0$ are unknown. Suppose we have a elicited a conjugate prior on (μ, σ^2), namely, $\mu \,|\, \sigma^2 \sim N(\mu_1, \tau_1\sigma^2)$ and $1/\sigma^2 \sim \text{Gamma}_{\text{rate}}(\alpha_1, \beta_1)$, and we would prefer to use a prior that is asymptotically uniformly weakly informative relative to this choice. As discussed in Evans and Moshonov (2006, 2007) it

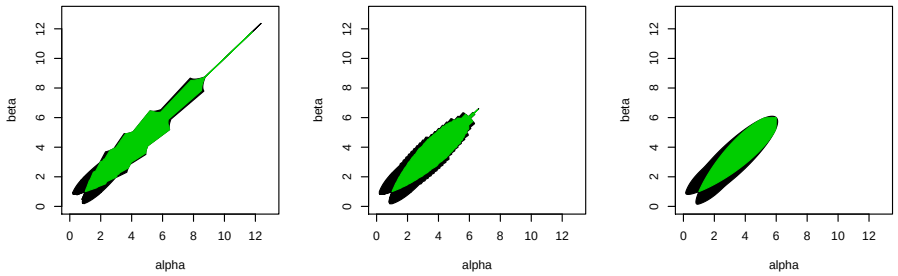

Figure 2: Plot of (α, β) corresponding to weakly informative priors at level $\gamma = 0.05$ (light and dark shading) and all (α, β) corresponding to uniformly weakly informative priors (light shading) for $n = 20, n = 100$ (middle), and $n = \infty$ (on the right).

seems that the most sensible way to check for prior-data conflict here is to first check the prior on $1/\sigma^2$, based on the prior predictive distribution of s^2, and, if no prior-data conflict is found at this stage, then check the prior on μ based on conditional prior predictive for $\bar{x}$ given s^2, as s^2 is ancillary for μ.

We have that $s^2 \,|\, \sigma^2 \sim \text{Gamma}_{\text{rate}}((n-1)/2, (n-1)/2\sigma^2)$ and so, as in Section 3.3, when $1/\sigma^2 \sim \text{Gamma}_{\text{rate}}(\alpha_i, \beta_i)$ the limiting prior predictive distribution of $1/s^2$ is $\text{Gamma}_{\text{rate}}(\alpha_i, \beta_i)$ as $n \to \infty$. Furthermore, when $T(x) = s^2$, then $J_T(x) = (4s^2/(n-1))^{-1/2}$. Therefore, the limiting value of (4) in this case is the same as that discussed in Section 3.3 and Theorem 4 applies to obtain a $\text{Gamma}_{\text{rate}}(\alpha_2, \beta_2)$ prior uniformly weakly informative relative to the $\text{Gamma}_{\text{rate}}(\alpha_1, \beta_1)$ prior.

If we consider s^2 as an arbitrary fixed value from its prior predictive distribution, then the conditional prior predictive distribution of $\bar{x} \,|\, s^2$ converges to the $N(\mu_1, \tau_1 s^2)$ distribution. Then by the results in Section 3.1 the $N(\mu_1, \tau_2\sigma^2)$ prior is asymptotically weakly informative at level γ relative to the $N(\mu_1, \tau_1\sigma^2)$ prior whenever $1 - G_1((\tau_2 s^2/\tau_1 s^2)G_1^{-1}(1-\gamma)) = 1 - G_1((\tau_2/\tau_1)G_1^{-1}(1-\gamma)) \leq \gamma$ and this occurs if and only if $\tau_2 \geq \tau_1$. Note that the dependence on the unknown value of s^2 disappears. Furthermore, the $N(\mu_1, \tau_2\sigma^2)$ prior is asymptotically uniformly weakly informative with respect to the $N(\mu_1, \tau_1\sigma^2)$ if and only if $\tau_2 \geq \tau_1$. We can determine an appropriate value for τ_2 by specifying $p \in [0, 1]$ and then setting $\tau_2 = \tau_1 G_1^{-1}(1 - \gamma + p\gamma)/G_1^{-1}(1-\gamma)$.

While this analysis is for a normal location-scale model, it is easy to see that the analysis for a general normal linear model will proceed along similar lines.

4.3 Weakly Informative Priors for Logistic Regression

Suppose we have a single binary valued response variable Y and k quantitative predictors $X_1, \ldots, X_k$, we observe $(Y, X_1, \ldots, X_k)$ at q settings of the predictor variables and have n_i observations at the i-th setting of the predictors. The logistic regression model then says that $Y_{ij} \sim \text{Bernoulli}(p_i)$ where $\log(p_i/(1-p_i)) = \beta_0 + \beta_1(x_{i1} - \bar{x}_{\cdot 1}) + \cdots + \beta_k(x_{ik} - \bar{x}_{\cdot k})$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, q$ and the β_i are unknown real values. For this model $T = (T_1, \ldots, T_q)$, with $T_i = Y_{i1} + \cdots + Y_{in_i}$, is a minimal sufficient statistic. For the base prior we

suppose that Π_1 is the product of independent priors on the β_i's and we consider the problem of finding a prior Π_2 that is weakly informative relative to Π_1. For example, we could take Π_1 to be a product of $N(0, \sigma_{1i}^2)$ priors and Π_2 to be a product of $N(0, \sigma_{2i}^2)$ priors and choose the σ_{2i}^2 so that weak informativity is obtained. Note that since T is discrete we can use (2) in our computations.

As we will see, it is not the case that choosing the σ_{2i}^2 very large relative to the σ_{1i}^2 will necessarily make Π_2 weakly informative relative to Π_1. In fact there is only a finite range of σ_{2i}^2 values where weak informativity will obtain.

While this can be demonstrated analytically, the argument is somewhat technical and it is perhaps easier to see this in an example. The following bioassay data are from Racine *et al.* (1986) and were also analyzed in Gelman *et al.* (2008). These data arise from an experiment where 20 animals were exposed to four doses of a toxin and the number of deaths recorded.

Dose (g/ml)	Number of animals n_i	Number of deaths t_i
0.422	5	0
0.744	5	1
0.948	5	3
2.069	5	5

Following Gelman *et al.* (2008) we took X_1 to be the variable formed by calculating the logarithm of dose and then standardizing to make the mean of X_1 equal to 0 and its standard deviation equal to 1/2. Gelman *et al.* (2008) placed independent Cauchy priors on the regression coefficients, namely, $\beta_0 \sim t_1(0, 10^2, 1)$ independent of $\beta_1 \sim t_1(0, 2.5^2, 1)$.

We consider four possible scenarios for the investigation of weak informativity at level $\gamma = 0.05$ and uniform weak informativity. In Figure 3 (a) we compare $\Pi_2 = N(0, \sigma_0^2) \times N(0, \sigma_1^2)$ priors with the prior $\Pi_1 = N(0, 10^2) \times N(0, 2.5^2)$. The entire region gives the (σ_0, σ_1) values corresponding to priors that are weakly informative at level $\gamma = 0.05$ while the lighter subregion gives the (σ_0, σ_1) values corresponding to priors that are uniformly weakly informative. Note that some of the irregularity in the plots is caused by the fact that the prior predictive distributions of T are discrete. The three remaining plots are similar where, in Figure 3(b) $\Pi_1 = t_1(0, 10^2, 1) \times t_1(0, 2.5^2, 1)$ and $\Pi_2 = t_1(0, \sigma_0^2, 1) \times t_1(0, \sigma_1^2, 1)$, in Figure 3(c) $\Pi_1 = N(0, 10^2) \times N(0, 2.5^2)$ and $\Pi_2 = t_1(0, \sigma_0^2, 1) \times t_1(0, \sigma_1^2, 1)$, and in Figure 3(d) $\Pi_1 = t_1(0, 10^2, 1) \times t_1(0, 2.5^2, 1)$ and $\Pi_2 = N(0, \sigma_0^2) \times N(0, \sigma_1^2)$. Note that these plots only depend on the data through the values of X_1.

We see clearly from these plots that increasing the scaling on any of the β_i does not necessarily lead to weak informativity and in fact inevitably destroys it. Furthermore, a smaller scaling on a parameter can lead to uniform weak informativity. These plots underscore how our intuition does not work very well with the logistic regression model as it is not clear how priors on the β_i ultimately translate to priors on the p_i. In fact it can be proven that, if we put independent priors on the β_i, fix all the scalings but one and let that one grow arbitrarily large, then the prior predictive distribution of T converges to a distribution degenerate two points, e.g., when the scaling on β_0 increases these

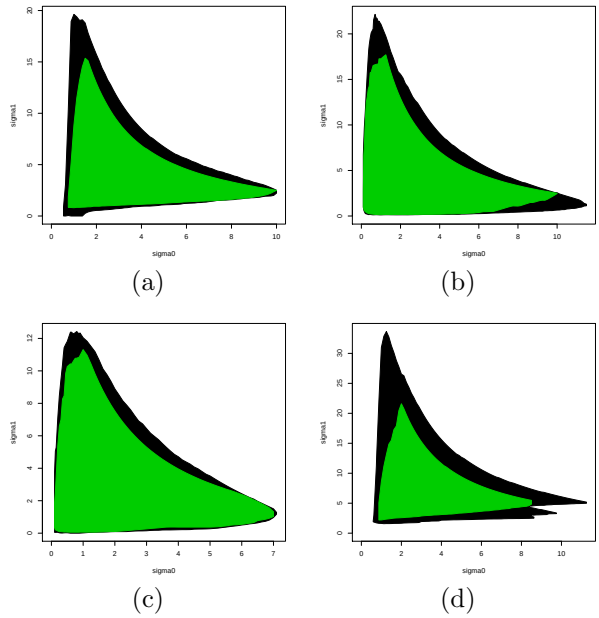

Figure 3: Weakly informative Π_2 priors relative to Π_1 at level 0.05 (light and dark shading) and uniformly weakly informative (light shading) for situations (a)-(d).

points are given by $\{\sum_{i=1}^{q} T_i = 0\} \cup \{\sum_{i=1}^{q} T_i = \sum_{i=1}^{q} n_i\}$, and this is definitely not desirable. This partially explains the results obtained.

Of some interest is how much reduction we actually get, via (5), when we employ a weakly informative prior. In Figure 4 we have plotted contours of the choices of (σ_0, σ_1) that give 0%, 25%, 50% and 75% reduction in prior-data conflicts for the case where $\Pi_2 = N(0, \sigma_0^2) \times N(0, \sigma_1^2)$ and $\Pi_1 = N(0, 10^2) \times N(0, 2.5^2)$ when $\gamma = 0.05$ (this corresponds to $x_\gamma = 0.0503$). Note that a substantial reduction can be obtained.

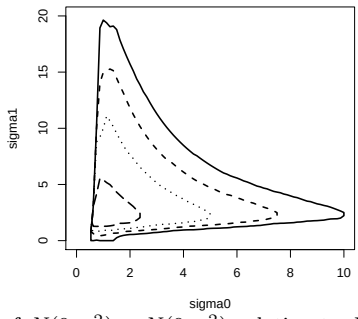

Figure 4: Reduction levels of $N(0, \sigma_0^2) \times N(0, \sigma_1^2)$ relative to $N(0, 10^2) \times N(0, 2.5^2)$ priors using (5) when $\gamma = 0.05$. The plotted reduction levels are 0% (solid line), 25% (dashed line), 50% (dotted line) and 75% (long dashed line).

We could consider fixing one of the scalings and seeing how much reduction

we obtain when varying the other. For example, when we fix $\sigma_0 = 2.5$ we find that the maximum reduction is obtained when σ_1 is close to 2.2628, while if we fix $\sigma_1 = 2.5$, then the maximum reduction is obtained when σ_0 is close to 0.875.

It makes sense in any application to check to see if any prior-data conflict exists with respect to the base prior. If there is no prior-data conflict this increases our confidence that the weakly informative prior is indeed putting less information into the analysis. This is assessed generally using (3) although (2) suffices in this example. When $\Pi_1 = N(0, 10^2) \times N(0, 2.5^2)$, then (2) equals 0.1073 and when $\Pi_1 = t_1(0, 10^2, 1) \times t_1(0, 2.5^2, 1)$ (the prior used in Gelman *et al.* (2008)), then (2) equals 0.1130, so in neither case is there any evidence of prior-data conflict.

5 Refinements Based Upon Ancillarity

The approach in Section 2 works whenever T is a complete minimal sufficient statistic. This is a consequence of Basu's Theorem as, in such a case, any ancillary is statistically independent of T and so conditioning on such an ancillary is irrelevant. When $U(T)$ is a meaningful ancillary, however, then the variation due to $U(T)$ is independent of θ and so should be removed from the P-value (3) when checking for prior-data conflict. Removing this variation is equivalent to conditioning on $U(T)$ and so we replace (3) by

$$M_T(m_T^*(t) \leq m_T^*(t_0) \,|\, U(T)), \tag{9}$$

i.e., we use the conditional prior predictive given the ancillary $U(T)$. To remove the maximal amount of ancillary variation we must have that $U(T)$ is a maximal ancillary. Therefore (4) becomes

$$M_{1T}(P_2(t_0 \,|\, U(T)) \leq x_\gamma \,|\, U(T)), \tag{10}$$

i.e., we have replaced M_{1T} by $M_{1T}(\cdot \,|\, U(T))$ and $P_2(t_0)$ by $P_2(t_0 \,|\, U(T)) = M_{2T}(m_{2T}^*(t) \leq m_{2T}^*(t_0) \,|\, U(T))$.

One problem with ancillaries is that multiple maximal ancillaries may exist. When ancillaries are used for frequentist inferences about θ via conditioning, this poses a problem because it is not clear which multiple ancillary to use and confidence regions depend on the maximal ancillary chosen. For checking for prior-data conflict via (9), however, this does not pose a problem. This is because we simply get different checks depending on which maximal ancillary we condition on. For example, if conditioning on maximal ancillary $U_1(T)$ does not lead to prior-data conflict, but conditioning on maximal ancillary $U_2(T)$ does, then we have evidence against no prior-data conflict existing.

Similarly, when we go to use (10), we can also simply look at the effect of each maximal ancillary on the analysis and make our assessment about Π_2 based on this. For example, we can use the maximum value of (10) over all maximal ancillaries to assess whether or not Π_2 is weakly informative relative to Π_1. When this maximum is small, we conclude that we have a small prior probability of

finding evidence against the null hypothesis of no prior-data conflict when using Π_2. We illustrate this via an example

Example

Suppose that we have a sample of n from the Multinomial$(1, (1-\theta)/6, (1+\theta)/6, (2-\theta)/6, (2+\theta)/6)$ distribution where $\theta \in [-1, 1]$ is unknown. Then the counts (f_1, f_2, f_3, f_4) constitute a minimal sufficient statistic and $U_1 = (f_1 + f_2, f_3 + f_4)$ is ancillary as is $U_2 = (f_1 + f_4, f_2 + f_3)$. Then $T = (f_1, f_2, f_3, f_4) \,|\, U_1$ is given by $f_1 \,|\, U_1 \sim$ Binomial$(f_1 + f_2, (1-\theta)/2)$ independent of $f_3 \,|\, U_1 \sim$ Binomial$(f_3 + f_4, (2-\theta)/4)$ giving

$$m_T(f_1, f_2, f_3, f_4 \,|\, U_1) = \binom{f_1 + f_2}{f_1}\binom{f_3 + f_4}{f_3} \times \int_{-1}^{1} \left(\frac{1-\theta}{2}\right)^{f_1} \left(\frac{1+\theta}{2}\right)^{f_2} \left(\frac{2-\theta}{4}\right)^{f_3} \left(\frac{2+\theta}{4}\right)^{f_4} \pi(\theta)\, d\theta.$$

We then have two 1-dimensional distributions $f_1 \,|\, U_1$ and $f_3 \,|\, U_1$ to use for checking for prior-data conflict. A similar result holds for the conditional distribution given U_2.

For example, suppose π is a Beta$(20, 20)$ distribution on $[-1, 1]$, so the prior concentrates about 0, and for a sample of $n = 18$ we have that $U_1 = f_1 + f_2 = 10$ and $U_2 = f_1 + f_4 = 8$. In Figure 5 we have plotted all the values of (α, β) that correspond to a Beta(α, β) prior that is weakly informative relative to the Beta$(20, 20)$ prior at level $\gamma = 0.05$ as well as those that are uniformly weakly informative. So for each such (α, β) we have that (10) is less than or equal to 0.05 for both $U = U_1$ and $U = U_2$.

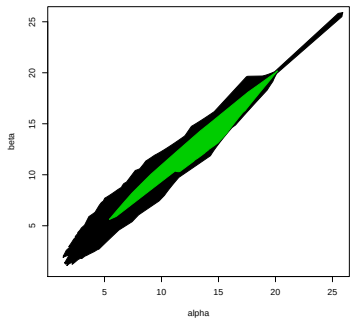

Figure 5: Plot of all (α, β) corresponding to weakly informative priors at level $\gamma = 0.05$ (light and dark shading) and uniformly weakly informative priors (light shading).

6 Conclusions

We have developed an approach to measuring the amount of information a prior puts into a statistical analysis relative to another base prior. This base prior can be considered as the prior that best reflects current information and our goal is to determine a prior that is weakly informative with respect to it. Our measure

is in terms of the prior predictive probability, using the base prior, of obtaining a prior-data conflict. This was applied in several examples where the approach is seen to give intuitively reasonable results.

As noted in several examples, however, we need to be careful when we conceive of a prior being weakly informative relative to another. Ultimately this concept needs to be made precise and we feel our definition is a reasonable proposal. The definition has intuitive support, in terms of avoiding prior-data conflicts, and provides a quantifiable criterion that can be used to select priors. This entails choosing a γ and using (5).

In any application we should still check for prior-data conflict for the base prior using (3). If prior-data conflict is found, a substitute prior that is weakly informative relative to the base prior can then be selected and a check made for prior-data conflict with respect to the new prior. In this way the new prior still incorporates some of the information from the base prior. If this new prior passes the checks for prior-data conflict, then the same theory can be applied to select a weakly informative prior relative to it, so that our inferences can be described as conservative.

We note that in several of the examples we have discussed, e.g., comparing normal priors, it is the case that a prior-data conflict with respect to the base prior can be avoided by a suitably chosen prior that is weakly informative relative to it. It doesn't seem possible, however, to prove that we can always find a weakly informative prior in a family that will avoid a prior-data conflict found for the base prior.

7 Appendix

Proof of Lemma 1 We have that $x_\gamma = \gamma$ since $P_1(t)$ has a continuous distribution under M_{1T}. Suppose $m^*_{iT}(t)$ has a point mass at r_0 when $t \sim M_{iT}$. The assumption $M_{iT}(m^*_{iT}(t) = r_0) > 0$ implies $(m^*_{iT})^{-1}\{r_0\} \neq \emptyset$. Then, pick $t_{r_0} \in (m^*_{iT})^{-1}\{r_0\}$ so that $m^*_{iT}(t_{r_0}) = r_0$ and let $\eta_i = P_i(t_{r_0})$. Then, $P_i(t)$ has point mass at η_i because $M_{iT}(P_i(t) = \eta_i) \geq M_{iT}(m^*_{iT}(t) = m^*_{iT}(t_{r_0})) = M_{iT}(m^*_{iT}(t) = r_0) > 0$. This is a contradiction and so $m^*_{iT}(t)$ has a continuous distribution when $t \sim M_{iT}$.

Let $r_\gamma = \sup\{r \in \mathcal{R} : M_{2T}(m^*_{2T}(t) \leq r) \leq \gamma\}$ where $\mathcal{R} = \overline{\{m^*_{2T}(t) : t \in \mathcal{T}\}}$ and $\mathcal{T}$ is the range space of T. Then, $M_{2T}(m^*_{2T}(t) \leq r_\gamma) = \gamma$ and $M_{2T}(m^*_{2T}(t) \leq r_\gamma + \epsilon) > \gamma$ for all $\epsilon > 0$. Thus, we have that $\{t : P_2(t) \leq \gamma\} = \{t : m^*_{2T}(t) \leq r_\gamma\}$, $M_{1T}(P_2(t) \leq \gamma) = M_{1T}(m^*_{2T}(t) \leq r_\gamma)$, and Π_2 is weakly informative at level γ relative to Π_1 if and only if $M_{1T}(m^*_{2T}(t) \leq r_\gamma) \leq \gamma$. The fact that $\{r_\gamma : \gamma \in [0,1]\} \subset \mathcal{R}$ implies the last statement.

Proof of Theorem 1 Suppose first that $\Sigma_1 \leq \Sigma_2$. We have that $n^{-1}I + \Sigma_1 \leq n^{-1}I + \Sigma_2$ and so $(n^{-1}I + \Sigma_1)^{-1} \geq (n^{-1}I + \Sigma_2)^{-1}$. This implies that (7) is less than γ and so the $N_k(\mu_0, \Sigma_2)$ prior is uniformly weakly informative relative to the $N_k(\mu_0, \Sigma_1)$ prior.

For the converse put $V_i = \{y : y'(n^{-1}I + \Sigma_i)^{-1}y \le 1\}$. If $V_1 \subset V_2$, then for $y \in R^k \backslash \{0\}$ there exists $c > 0$ such that $c^2 y'(n^{-1}I + \Sigma_1)^{-1}y = 1$ which implies $cy \in V_2$ and so $c^2 y'(n^{-1}I + \Sigma_2)^{-1}y \le 1$. This implies that $y'(n^{-1}I + \Sigma_1)^{-1}y \ge y'(n^{-1}I + \Sigma_2)^{-1}y$ and so $\Sigma_1 \le \Sigma_2$ and the result follows. If $V_2 \subset V_1$ then the same reasoning says that $\Sigma_2 \le \Sigma_1$ and (7) would be greater than γ if $\Sigma_2 < \Sigma_1$.

So we need only consider the case where $V_1 \cap V_2^c$, $V_1^c \cap V_2$ both have positive volume, i.e., we are supposing that neither $\Sigma_2 - \Sigma_1$ nor $\Sigma_1 - \Sigma_2$ is positive semidefinite and then will obtain a contradiction. Let $\delta = \inf\{y'(n^{-1}I + \Sigma_1)^{-1}y : y \in V_1 \cap \partial V_2\}$ and note that $\delta < 1$, since $V_1^o \cap \partial V_2 \ne \phi$, i.e., there are points in the interior of V_1 on the boundary of V_2. Now put $V_0 = \{y \in V_1 \cap V_2^c : y'(n^{-1}I + \Sigma_1)^{-1}y \le (1+\delta)/2\}$ and note that V_0 has positive volume.

Let $Y \sim N_k(0, n^{-1}I + \Sigma_1)$ and $\tau_\gamma^2 = G_k^{-1}(1-\gamma)$. Then $M_{1T}(P_1(t) \le \gamma) = P(Y'(n^{-1}I + \Sigma_1)^{-1}Y \ge \tau_\gamma^2) = P(Y \notin \tau_\gamma V_1) = 1 - P_Y(\tau_\gamma(V_1 \cap V_2) \cup \tau_\gamma(V_1 \cap V_2^c))$ while $M_{1T}(P_2(t) \le \gamma) = P(Y'(n^{-1}I + \Sigma_2)^{-1}Y \ge \tau_\gamma^2) = P(Y \notin \tau_\gamma V_2) = 1 - P_Y(\tau_\gamma(V_1 \cap V_2) \cup \tau_\gamma(V_1^c \cap V_2))$. Since $\gamma = M_{1T}(P_1(t) \le \gamma)$, we need only show that $P_Y(\tau_\gamma(V_1 \cap V_2^c)) > P_Y(\tau_\gamma(V_1^c \cap V_2))$ for all γ sufficiently small, to establish the result.

Let $f(x) = k_1 e^{-x/2}$ be such that $f(y'(n^{-1}I + \Sigma_1)^{-1}y)$ is the density of Y. Then $P_Y(\tau_\gamma(V_1^c \cap V_2)) = \int_{\tau_\gamma(V_1^c \cap V_2)} f(y'(n^{-1}I + \Sigma_1)^{-1}y)\, dy \le f(\tau_\gamma^2 y_*'(n^{-1}I + \Sigma_1)^{-1}y_*) \operatorname{Vol}((V_1^c \cap V_2))\tau_\gamma^k$ where $y_* = \arg\min\{y'(n^{-1}I + \Sigma_1)^{-1}y : y \in V_1^c \cap V_2\}$. Note it is clear that $y_* \in \partial V_1$ and so $y_*'(n^{-1}I + \Sigma_1)^{-1}y_* = 1$ and $f(\tau_\gamma^2 y_*'(n^{-1}I + \Sigma_1)^{-1}y_*) = k_1 e^{-\tau_\gamma^2/2}$. Also, $P_Y(\tau_\gamma(V_1 \cap V_2^c)) \ge P_Y(\tau_\gamma V_0) = \int_{\tau_\gamma V_0} f(y'(n^{-1}I + \Sigma_1)^{-1}y)\, dy \ge f(\tau_\gamma^2(1+\delta)/2) \operatorname{Vol}(V_0)\tau_\gamma^k$ where $f(\tau_\gamma^2(1+\delta)/2) = k_1 e^{-\tau_\gamma^2(1+\delta)/4}$. Therefore, as $\gamma \to 0$,

$$\frac{P_Y(\tau_\gamma(V_1 \cap V_2^c))}{P_Y(\tau_\gamma(V_1^c \cap V_2))} \ge e^{\tau_\gamma^2(1-\delta)/4} \frac{\operatorname{Vol}((V_1^c \cap V_2))}{\operatorname{Vol}(V_0)} \to \infty$$

since $\tau_\gamma = (G_k^{-1}(1-\gamma))^{1/2} \to \infty$ as $\gamma \to 0$ and $0 < \delta < 1$.

Proof of Theorem 2 First note that we can use (2) instead of (3) in this case as $J_T(x)$ is constant in this case. We assume without loss of generality that $\mu_0 = 0$.

We first establish several useful technical results. If Π_i is a probability distribution that is unimodal and symmetric about 0, and ϕ_ν denotes a $N(0, \nu)$ density, we have that $m_{iT}(t) = \int_R \phi_\nu(t - \mu)\, \Pi_i(d\mu)$ is unimodal and symmetric about 0. We have the following result.

Lemma A1. If T is a minimal sufficient statistic, $J_T(x)$ is constant in x, Π_1 and Π_2 are unimodal and symmetric about 0, the $P_i(t)$ have continuous distributions when $t \sim M_{iT}, m_{1T}(0) > m_{2T}(0)$, and $m_{1T}(t) = m_{2T}(t)$ has a unique solution for $t > 0$, then Π_2 is uniformly weakly informative relative to Π_1.

Proof: By the unimodality and symmetry of m_{iT}, we have $P_i(t) = M_{iT}(m_{iT}(u) \le m_{iT}(t)) = M_{iT}(|u| \ge |t|)$. We show $M_{1T}(|t| \ge t_0) \le M_{2T}(|t| \ge t_0)$ for all $t_0 > 0$ because it is equivalent to Π_2 being uniformly weakly informative relative to Π_1 by Lemma 1. Let t_s be the solution of $m_{1T}(t) = m_{2T}(t)$ on $(0, \infty)$.

From the unique solution assumption, $m_{1T}(t) > m_{2T}(t)$ for $t \in (0, t_s)$ and $m_{1T}(t) < m_{2T}(t)$ for $t > t_s$. For $0 \le t_0 < t_s$, $M_{1T}(|t| \ge t_0) = 2\int_{t_0}^{\infty} m_{1T}(t)\,dt = 1 - 2\int_0^{t_0} m_{1T}(t)\,dt \le 1 - 2\int_0^{t_0} m_{2T}(t)\,dt = 2\int_{t_0}^{\infty} m_{2T}(t)\,dt = M_{2T}(|t| \ge t_0)$ and for $t_0 \ge t_s$, $M_{1T}(|t| \ge t_0) = 2\int_{t_0}^{\infty} m_{1T}(t)\,dt \le 2\int_{t_0}^{\infty} m_{2T}(t)\,dt = M_{2T}(|t| \ge t_0)$. Thus, we are done.

We can apply Lemma A1 to comparing normal and t priors when sampling from a normal.

Lemma A2. Suppose we have a sample of n from a location normal model, Π_1 is a $N(0, \sigma_1^2)$ prior and Π_2 is a $t_1(0, \sigma_2^2, \lambda)$ prior. If $m_{1T}(0) > m_{2T}(0)$, then Π_2 is uniformly weakly informative relative to Π_1.

Proof: We have that $m_{1T} = \phi_{1/n+\sigma_1^2}$ and, using the representation of the $t(\lambda)$ distribution as a gamma mixture of normals, we can write $m_{2T}(t) = \int_0^{\infty} \phi_{1/n+\sigma_2^2/u}(t) k_\lambda(u) du$ where k_λ is the density of $\text{Gamma}_{\text{rate}}(\lambda/2, \lambda/2)$ distribution. By the symmetry of ϕ_v, m_{2T} is symmetric. Also $\phi_v(t_1) > \phi_v(t_2)$ for $0 \le t_1 < t_2$ and so $m_{2T}(t_1) = \int \phi_{1/n+\sigma_2^2/u}(t_1) k_\lambda(u)\,du \ge \int \phi_{1/n+\sigma_2^2/u}(t_2) k_\lambda(u)\,du = m_{2T}(t_2)$. Thus, m_{2T} is decreasing on $(0, \infty)$, i.e., m_{2T} is unimodal. To show that $m_{2T}(t)$ is log-convex with respect to t^2 we prove that $(d^2/d(t^2)^2) \log m_{2T}(t) \ge 0$. Note that $(d/d(t^2))\phi_v(t) = (d/d(t^2))[(2\pi v)^{-1/2} \exp\{-t^2/2v\}] = -\phi_v(t)/2v$,

$$\begin{aligned}
\frac{dm_{2T}(t)}{dt^2} &= -\int_0^{\infty} \frac{\phi_{1/n+\sigma_2^2/u}(t)}{2(1/n + \sigma_2^2/u)} k_\lambda(u)\,du, \\
\frac{d^2 \log m_{2T}(t)}{d(t^2)^2} &= \frac{1}{m_{2T}(t)} \int_0^{\infty} \frac{\phi_{1/n+\sigma_2^2/u}(t)}{[2(1/n + \sigma_2^2/u)]^2} k_\lambda(u)\,du - \\
&\qquad \frac{1}{m_{2T}(t)^2} \Big(\int_0^{\infty} \frac{\phi_{1/n+\sigma_2^2/u}(t)}{2(1/n + \sigma_2^2/u)} k_\lambda(u)\,du \Big)^2
\end{aligned}$$

and so $d^2 \log m_{2T}(t)/d(t^2)^2 = \text{Var}_V([2(1/n + \sigma_2^2/V)]^{-1}) \ge 0$, where V is the random variable having density $\phi_{1/n+\sigma_2^2/v}(t) k_\lambda(v)/m_{2T}(t)$. Thus, $m_{2T}(t)$ is log-convex in t^2.

The functions $m_{1T}(t)$ and $m_{2T}(t)$ meet in at most two points on $(0, \infty)$ because $\log m_{1T}(t)$ is linear in t^2 and $\log m_{2T}(t)$ is convex in t^2. Also $m_{1T}(t)$ and $m_{2T}(t)$ share at least one point on $(0, \infty)$ because $m_{1T}(0) > m_{2T}(0)$, and the following shows that $m_{1T}(t) < m_{2T}(t)$ for all large t. Note first that if $u \ge \sigma_2^2/2\sigma_1^2$, then $(1/n+\sigma_1^2)/(1/n+\sigma_2^2/u) \ge 1/2$ and $t^2/(u/n+\sigma_2^2) \ge (2\sigma_1^2/\sigma_2^2)t^2/(1/n+2\sigma_1^2)$. Then,

$$\begin{aligned}
\frac{m_{2T}(t)}{m_{1T}(t)} \ge \int_{\sigma_2^2/2\sigma_1^2}^{\infty} & \frac{(\lambda/2)^{\lambda/2}}{\Gamma(\lambda/2)} \frac{(2\pi(1/n + \sigma_2^2/u))^{-1/2}}{(2\pi(1/n + \sigma_1^2))^{-1/2}} u^{\lambda/2-1} \times \\
& \frac{\exp\{-(u/2)(\lambda + t^2/(u/n + \sigma_2^2))\}}{\exp\{-(1/2)t^2/(1/n + \sigma_1^2)\}}\, du
\end{aligned}$$

$$\geq \frac{(\lambda/2)^{\lambda/2}}{\Gamma(\lambda/2)}\frac{1}{2^{1/2}}\Big(\frac{\sigma_2^2}{2\sigma_1^2}\Big)^{\lambda/2-1}\times$$
$$\int_{\sigma_2^2/2\sigma_1^2}^{\infty}\frac{\exp\{-(u/2)(\lambda+(2\sigma_1^2/\sigma_2^2)t^2/(1/n+2\sigma_1^2))\}}{\exp\{-(1/2)t^2/(1/n+\sigma_1^2)\}}\,du$$
$$= \frac{(\lambda/2)^{\lambda/2}}{\Gamma(\lambda/2)}\frac{1}{2^{1/2}}\Big(\frac{\sigma_2^2}{2\sigma_1^2}\Big)^{\lambda/2-1}\exp\{-(1/2)(\sigma_2^2/2\sigma_1^2)\lambda\}\times$$
$$\frac{\exp\{(1/2)t^2((1/n+\sigma_1^2)^{-1}-(1/n+2\sigma_1^2)^{-1})\}}{2^{-1}(\lambda+(2\sigma_1^2/\sigma_2^2)t^2/(1/n+2\sigma_1^2))}\to\infty$$

as $t^2\to\infty$.

The above conditions together imply that $m_{1T}(t)$ and $m_{2T}(t)$ meet in exactly one point on $(0,\infty)$. Therefore, Π_2 is uniformly weakly informative relative to Π_1 by Lemma A1.

Since $\int_0^\infty(1/n+\sigma^2/u)^{-1/2}k_\lambda(u)\,du$ is strictly decreasing in σ^2, we see that $m_{1T}(0)=(2\pi(1/n+\sigma_1^2))^{-1/2}\geq m_{2T}(0)=(2\pi)^{-1/2}\int_0^\infty(1/n+\sigma_2^2/u)^{-1/2}k_\lambda(u)\,du$ is equivalent to $\sigma_2\geq\sigma_{0n}$ where σ_{0n} satisfies $(1/n+\sigma_1^2)^{-1/2}=\int_0^\infty(1/n+\sigma_{0n}^2/u)^{-1/2}k_\lambda(u)\,du$. This proves the first part of Theorem 2.

We also need the following results for the remaining parts of Theorem 2.

Lemma A3. (i) σ_{0n}^2/σ_1^2 increases as $n\sigma_1^2\to\infty$, (ii) $\sigma_{0n}^2/\sigma_1^2\to(2/\lambda)\Gamma^2((\lambda+1)/2)/\Gamma^2(\lambda/2)$ as $n\sigma_1^2\to\infty$.

Proof: (i) We have $n^{-1/2}(1/n+\sigma_1^2)^{-1/2}=n^{-1/2}\int_0^\infty(1/n+\sigma_{0n}^2/u)^{-1/2}k_\lambda(u)\,du$ and putting $\alpha=n\sigma_1^2,\beta=n\sigma_{0n}^2$ we can write this as

$$(1+\alpha)^{-1/2}=\int_0^\infty(1+\beta/u)^{-1/2}k_\lambda(u)\,du. \tag{A1}$$

Differentiating both sides of (A1) with respect to α we have $(1+\alpha)^{-3/2}=\int_0^\infty(1+\beta/u)^{-3/2}u^{-1}k_\lambda(u)\,du\,(d\beta/d\alpha)$. If we let $U\sim\text{Gamma}_{\text{rate}}(\lambda/2,\lambda/2)$, then this integral can be written as the expectation

$$E((1+\beta/U)^{-3/2}U^{-1})=E((1+\beta/U)^{-3/2}(\beta/U+1-1)/\beta)$$
$$=\beta^{-1}E((1+\beta/U)^{-1/2})-\beta^{-1}E((1+\beta/U)^{-3/2})$$
$$\leq\beta^{-1}E((1+\beta/U)^{-1/2})-\beta^{-1}\{E((1+\beta/U)^{-1/2})\}^3$$
$$=\beta^{-1}(1+\alpha)^{-1/2}-\beta^{-1}(1+\alpha)^{-3/2}=(1+\alpha)^{-3/2}(\alpha/\beta)$$

where the inequality follows via Jensen's inequality. Hence, $d\beta/d\alpha=(1+\alpha)^{-3/2}/E((1+\beta/U)^{-3/2}U^{-1})\geq\beta/\alpha$ and so β/α is an increasing function of α because $d(\beta/\alpha)/d\alpha=\alpha^{-1}(d\beta/d\alpha)-\beta/\alpha^2\geq 0$. This proves $\sigma_{0n}^2/\sigma_1^2=n\sigma_{0n}^2/n\sigma_1^2=\beta/\alpha$ increases as $\alpha=n\sigma_1^2\to\infty$.

(ii) It is easy to check that $\beta=0$ when $\alpha=0$ and $\beta>0$ for $\alpha>0$. Let α_0,β_0 be a pair satisfying $\alpha_0>0$ and (A1). Then, $\beta/\alpha\geq\beta_0/\alpha_0>0$ for $\alpha>\alpha_0$ and

$\beta \to \infty$ as $\alpha \to \infty$. Therefore,

$$\lim_{\alpha\to\infty}\left(\frac{\beta}{\alpha}\right)^{1/2} = \lim_{\alpha\to\infty}\frac{\sqrt{\beta}}{\sqrt{1+\alpha}} = \lim_{\alpha\to\infty} E\left(\frac{\sqrt{\beta}}{\sqrt{1+\beta/U}}\right)$$
$$= \lim_{\beta\to\infty} E\left(\frac{\sqrt{\beta}}{\sqrt{1+\beta/U}}\right) = E(\sqrt{U})$$
$$= \int_0^\infty \sqrt{u}\, k_\lambda(u)\, du = (2/\lambda)^{1/2}\Gamma((\lambda+1)/2)/\Gamma(\lambda/2)$$

and this proves (ii).

Lemma A4. Suppose we have a sample of n from a location normal model, Π_1 is a $N(0,\sigma_1^2)$ prior and Π_2 is a $t_1(0,\sigma_2^2,\lambda)$ prior. Then Π_2 is asymptotically uniformly weakly informative relative to Π_1 if and only if $\sigma_2^2/\sigma_1^2 \geq (2/\lambda)\Gamma^2((\lambda+1)/2)/\Gamma^2(\lambda/2)$.

Proof: Suppose that $\sigma_2^2/\sigma_1^2 \geq (2/\lambda)\Gamma^2((\lambda+1)/2)/\Gamma^2(\lambda/2)$. Then by Lemma A3 $\sigma_2^2/\sigma_1^2 \geq \sigma_{0n}^2/\sigma_1^2$ for all n and so Π_2 is uniformly weakly informative with respect to Π_1 for all n. So (4) is bounded above by γ for all n and so the limiting value of (4) is also bounded above by γ. This establishes that Π_2 is asymptotically uniformly weakly informative relative to Π_1.

Suppose now that $\sigma_2^2/\sigma_1^2 < (2/\lambda)\Gamma^2((\lambda+1)/2)/\Gamma^2(\lambda/2)$. Note that $m_{1T}(t) = \lim_{n\to\infty} m_{1T,n}(t) = (2\pi\sigma_1^2)^{-1/2}\exp(-t^2/(2\sigma_1^2))$ and $m_{2T}(t) = \lim_{n\to\infty} m_{2T,n}(t)$ $= \Gamma((\lambda+1)/2)/(\Gamma(\lambda/2)\sqrt{\pi\lambda\sigma_2^2})(1+x^2/(\sigma_2^2\lambda))^{-(\lambda+1)/2}$. Therefore, $m_{1T}(0) = 1/\sqrt{2\pi\sigma_1^2} < \Gamma((\lambda+1)/2)/\Gamma(\lambda/2)\sqrt{\pi\lambda\sigma_2^2} = m_{2T}(0)$. Let $B = \{t : m_{2T}(t) > m_{1T}(0)\}$ and $\gamma = M_{2T}(B^c)$. Then, $m_{1T}(t) \leq m_{1T}(0) \leq m_{2T}(t)$ on B and $M_{1T}(P_2(t) \leq \gamma) = M_{1T}(B^c) = 1 - M_{1T}(B) = 1 - \int_B m_{1T}(t)\,dt \geq 1 - \int_B m_{1T}(0)\,dt$ $\geq 1 - \int_B m_{2T}(t)dt = M_{2T}(B^c) = \gamma$. Hence, Π_2 is not weakly informative relative to Π_1 at level γ. Therefore, $\sigma_2^2/\sigma_1^2 \geq (2/\lambda)\Gamma^2((\lambda+1)/2)/\Gamma^2(\lambda/2)$.

It is now immediate that $\sup_{\gamma\in[0,1]} G_1^{-1}(1-\gamma)/H_{1,\lambda}^{-1}(1-\gamma) = (2/\lambda)\Gamma^2((\lambda+1)/2)/\Gamma^2(\lambda/2)$ and the proof of Theorem 2 is complete.

Proof of Theorem 3 Since the minimal sufficient statistic $T(x) = \bar{x}$ is linear there is no volume distortion and we can use (2) instead of (3). The limiting prior predictive distribution of $T(x) = \bar{x}$ under Π_1 is $N(\mu_0, \Sigma_1)$ and under Π_2 it is $t_k(\mu_0, \Sigma_2, \lambda)$. It is easy to check that $U_1 = (T-\mu_0)'\Sigma_1^{-1}(T-\mu_0) \sim \chi^2(k)$ when $T \sim \Pi_1$ and $U_2 = (T-\mu_0)'\Sigma_2^{-1}(T-\mu_0) \sim kF_{k,\lambda}$ when $T \sim \Pi_2$. This implies that, $P_{2,n}(t_0)$ converges to $P_2(t_0) = \Pi_2(\pi_2(t) \leq \pi_2(t_0)) = 1 - H_{k,\lambda}((t_0-\mu_0)'\Sigma_2^{-1}(t_0-\mu_0)/k)$ where $H_{k,\lambda}$ is the distribution function of an $F_{k,\lambda}$ distribution. Further, we have that (4) converges to $\Pi_1(P_2(t) \leq \gamma)$.

Let $V_i = \{u \in R^k : u'\Sigma_i^{-1}u < 1\}$ for $i = 1, 2$. By the continuity of $\Pi_2(\pi_2(t) \leq r)$ as a function of r, and the continuity of $\pi_2(t)$, there exists t_0 such that $P_2(t) \leq \gamma$ if and only if $\pi_2(t) \leq \pi_2(t_0)$. Hence, Π_2 is asymptotically uniformly weakly informative relative to Π_1 if and only if $\Pi_1(\pi_2(t) \leq \pi_2(t_0)) \leq \Pi_2(\pi_2(t) \leq \pi_2(t_0))$ for all $t_0 \in R^k$ by Lemma 1. Since $\pi_2(t)$ is decreasing in $u_2 = U_2(t)$, the set

$\{\pi_2(t) \le \pi_2(t_0)\} = \{u_2(t) \ge u_2(t_0)\} = \mu_0 + u_2(t_0)V_2^c$. So we must prove that $\Pi_1(\mu_0 + r^{1/2}V_2^c) \le \Pi_2(\mu_0 + r^{1/2}V_2^c)$ for all $r \ge 0$.

The positive semidefiniteness of $\Sigma_2 - \tau_\lambda^2\Sigma_1$ implies that $\Sigma_1^{-1}/\tau_\lambda^2 - \Sigma_2^{-1}$ is positive semidefinite. Then, for $u \in V_2^c$, that is, $u'\Sigma_2^{-1}u \ge 1$, we have $u'\Sigma_1^{-1}u = \tau_\lambda^2 \cdot u'(\Sigma_1^{-1}/\tau_\lambda^2)u \ge \tau_\lambda^2 u'\Sigma_2^{-1}u \ge \tau_\lambda^2$. Thus, $V_2^c \subset \tau_\lambda V_1^c$.

Now we prove a stronger inequality $\Pi_1(\mu_0 + r^{1/2}\tau_\lambda V_1^c) \le \Pi_2(\mu_0 + r^{1/2}V_2^c)$ for all $r \ge 0$. Note that

$$\Pi_1(\mu_0 + r^{1/2}\tau_\lambda V_1^c) = \Pi_1(u_1(t) \ge r\tau_\lambda^2) = \int_{r\tau_\lambda^2}^\infty \frac{2^{-k/2}}{\Gamma(k/2)} u^{k/2-1}e^{-u/2}\,du,$$

$$\begin{aligned}\Pi_2(\mu_0 + r^{1/2}V_2^c) &= \Pi_2(u_2(t) \ge r)\\ &= \int_{r/k}^\infty \frac{\Gamma((k+\lambda)/2)}{\Gamma(k/2)\Gamma(\lambda/2)}\Big(\frac{k}{\lambda}\Big)^{k/2} u^{k/2-1}\Big(1 + \frac{ku}{\lambda}\Big)^{-(k+\lambda)/2}\,du.\end{aligned}$$

and set $f(r) = \Pi_2(\mu_0 + r^{1/2}V_2^c) - \Pi_1(\mu_0 + r^{1/2}\tau_\lambda V_1^c)$. Then, $f(0) = 0$ and

$$\begin{aligned}\frac{df(r)}{dr} &= \frac{2^{-k/2}}{\Gamma(k/2)}(r\tau_\lambda^2)^{k/2-1}e^{-r\tau_\lambda^2/2}\tau_\lambda^2 -\\ &\qquad \frac{\Gamma((k+\lambda)/2)}{\Gamma(k/2)\Gamma(\lambda/2)}\Big(\frac{k}{\lambda}\Big)^{k/2}\Big(\frac{r}{k}\Big)^{k/2-1}\Big(1 + r/\lambda\Big)^{-(k+\lambda)/2}\frac{1}{k}\\ &= \frac{(\tau_\lambda^2/2)^{k/2}}{\Gamma(k/2)} r^{k/2-1}e^{-r\tau_\lambda^2/2} - \frac{\Gamma((k+\lambda)/2)}{\Gamma(k/2)\Gamma(\lambda/2)}\lambda^{-k/2}r^{k/2-1}\Big(1 + r/\lambda\Big)^{-(k+\lambda)/2}\\ &= p_1 - p_2.\end{aligned}$$

Note that $p_1 - p_2 \ge 0$ is equivalent to $p_1/p_2 \ge 1$. Further recalling the definition of τ_λ^2 from the statement of the theorem,

$$\begin{aligned}\frac{p_1}{p_2} &= \frac{\tau_\lambda^k\Gamma(\lambda/2)}{\Gamma((k+\lambda)/2)}\Big(\frac{\lambda}{2}\Big)^{k/2}\Big(1 + r/\lambda\Big)^{(k+\lambda)/2}\exp(-r\tau_\lambda^2/2)\\ &= \Big(1 + r/\lambda\Big)^{(k+\lambda)/2}\exp(-r\tau_\lambda^2/2) \ge 1.\end{aligned}$$

The logarithm of p_1/p_2 given by $\log(p_1/p_2) = -r\tau_\lambda^2/2 + ((k+\lambda)/2)\log(1+r/\lambda)$ is concave as a function of $r > 0$. Hence, $\log(p_1/p_2) = 0$ has exactly two solutions $r = 0$ and $r = r_s$. Because of its concavity, the function $\log(p_1/p_2)$ is positive on $(0, r_s)$ and negative on (r_s, ∞). This implies that $f(r)$ is increasing on $(0, r_s)$ and decreasing on (r_s, ∞). Since $f(0) = 0$ and $\lim_{r\to\infty} f(r) = 0$, the function f is nonnegative, that is, $f(r) \ge 0$ for all $r \ge 0$. Thus, $\Pi_1(\mu_0 + r^{1/2}V_2^c) \le \Pi_1(\mu_0 + r^{1/2}\tau_\lambda V_1^c) \le \Pi_2(\mu_0 + r^{1/2}V_2^c)$ for all $r \ge 0$.

Proof of Theorem 4 Let $x_c^{-1} = \beta_1/(\alpha_1 + 1/2) = \beta_2(\alpha_2 + 1/2)$. For $i = 1, 2$, let $t_i(t_0) = 1/(x_c r_i(t_0))$ be the two solutions of $m_{2T}^*(t_i) = m_{2T}^*(t_0)$ (one of the t_i equals t_0) so $0 < r_1 \le 1 \le r_2$. Note that $r_2(t_0) = 1$ if and only if $t_0 = x_c^{-1}$ and then $r_1(t_0) = 1$ as well. Then, $\log(r_1/r_2) = r_1 - r_2$ and $dr_1/dr_2 = (r_2 - 1)r_1/[(r_1 - 1)r_2]$. Now $\{t : m_{2T}^*(t) \le m_{2T}^*(t_0)\} = \{t : 1/t \le x_c r_1(t_0) \text{ or } 1/t \ge$

$x_c r_2(t_0)\}$. By Lemma 1 we have that uniform weak informativity is equivalent to $M_{1T}(m^*_{2T}(t) \le m^*_{2T}(t_0)) \le M_{2T}(m^*_{2T}(t) \le m^*_{2T}(t_0))$ for all t_0 and so we must prove that $M_{1T}(t \notin (t_2(t_0), t_1(t_0))) = M_{1T}(1/t \le x_c r_1(t_0) \text{ or } 1/t \ge x_c r_2(t_0)) = 1 - M_{1T}(x_c r_1(t_0) \le 1/t \le x_c r_2(t_0)) \le 1 - M_{2T}(x_c r_1(t_0) \le 1/t \le x_c r_2(t_0))$ for all t_0. Since r_1 is implicitly a function of r_2, it is equivalent to prove that $M_{1T}(x_c r_1 \le 1/t \le x_c r_2) - M_{2T}(x_c r_1 \le 1/t \le x_c r_2) \ge 0$ for all $r_2 \ge 1$. Using $(r_1/r_2)^\alpha = \exp(\alpha(r_1 - r_2))$, we have that the derivatives of the two terms are given by

$$\begin{aligned} p_1 &= \frac{d}{dr_2} \int_{x_c r_1}^{x_c r_2} c_1 u^{\alpha_1 - 1} e^{-\beta_1 u} du = c_1 (x_c r_2)^{\alpha_1 - 1} e^{-\beta_1 x_c r_2} x_c - \\ &\qquad c_1 (x_c r_1)^{\alpha_1 - 1} e^{-\beta_1 x_c r_1} x_c \frac{dr_1}{dr_2} \\ &= c_1 x_c^{\alpha_1} r_2^{\alpha_1 - 1} e^{-\beta_1 x_c r_2} \Big(1 - \frac{r_2 - 1}{r_1 - 1} \exp((r_2 - r_1)(\beta_1 x_c - \alpha_1))\Big), \\ p_2 &= c_2 x_c^{\alpha_2} r_2^{\alpha_2 - 1} e^{-\beta_2 x_c r_2} \Big(1 - \frac{r_2 - 1}{r_1 - 1} \exp((r_2 - r_1)(\beta_2 x_c - \alpha_2))\Big) \end{aligned}$$

where $c_i = \beta_i^{\alpha_i}/\Gamma(\alpha_i)$. Then, recalling the definition of x_c, we have that the ratio $p_1/p_2 = (c_1/c_2) x_c^{\alpha_1 - \alpha_2} r_2^{\alpha_1 - \alpha_2} e^{(\beta_2 - \beta_1) x_c r_2} = (c_1/c_2) x_c^{\alpha_1 - \alpha_2} (r_2 e^{-r_2})^{\alpha_1 - \alpha_2}$ strictly decreases as r_2 increases from 1 to ∞ when $\alpha_1 > \alpha_2$ because $\alpha_1 - \alpha_2 = (\beta_1 - \beta_2) x_c \ge 0$, and is identically 1 when $\alpha_1 = \alpha_2$. Suppose then that $\alpha_1 > \alpha_2$ so there is at most one r_2 value where $p_1 = p_2$ and the derivative is 0. If $(p_1/p_2)|_{r_2=1} < 1$, then $p_1 - p_2 < 0$ for all $r_2 \ge 1$ and $M_{1T}(x_c r_1 \le 1/t \le x_c r_2) - M_{2T}(x_c r_1 \le 1/t \le x_c r_2)$ strictly decreases from 0. This cannot hold because $M_{1T}(x_c r_1 \le 1/t \le x_c r_2) - M_{2T}(x_c r_1 \le 1/t \le x_c r_2) \to 0$ as $r_2 \to \infty$. Hence, $(p_1/p_2)|_{r_2=1} \ge 1$ and $M_{1T}(x_c r_1 \le 1/t \le x_c r_2) - M_{2T}(x_c r_1 \le 1/t \le x_c r_2)$ increases from 0 near $r_2 = 1$ and decreases to 0 as $r_2 \to \infty$. Therefore, $M_{1T}(x_c r_1 \le 1/t \le x_c r_2) - M_{2T}(x_c r_1 \le 1/t \le x_c r_2)$ goes up from 0 and down to 0 as r_2 increases from 1 to ∞, we have $M_{1T}(x_c r_1 \le 1/t \le x_c r_2) - M_{2T}(x_c r_1 \le 1/t \le x_c r_2) \ge 0$ for all $r_2 \ge 1$.

References

Bernardo, J.-M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41(2):113–147. With discussion.

Evans, M. and Jang, G. H. (2008). Invariant P-values for model checking and checking for prior-data conflict. Technical Report No. 0803, Department of Statistics, University of Toronto.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Anal.*, 1(4):893–914.

Evans, M. and Moshonov, H. (2007). Checking for prior-data conflict with hierarchically specified priors. In Upadhyay, A., Singh, U., and Dey, D.,

editors, *Bayesian Statistics and its Applications*, pages 145–159. Anamaya Publishers, New Delhi.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.*, 1(3):515–533.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.*, 90(431):928–934.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005.

Racine, A., Grieve, A. P., Flühler, H., and Smith, A. F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry. *J. Roy. Statist. Soc. Ser. C*, 35(2):93–150. With discussion.